## Team Name:

| 7b |

## Team Members and email addresses:

- Kunal Karnik (kunalkarnik95@gmail.com)
- Michael Neises (neisesmichael@gmail.com)
- Brian McClannahan (b523m844@ku.edu)
- Omar Alzubbi (o299a438@ku.edu)
- Quentin Wiley (qtwiley95@gmail.com)
- Ryan Feehan (rfeehan93@gmail.com)

## Team Meeting time:

- Friday 4:00 – 6:00p

## Lab Meeting time:

- Friday 4:00 – 6:00p

## Contact:

- Kunal Karnik

## Project Sponsor:

- Black and Veatch

## Project Description (150-250 words)

**DataDriven** is a data driven merit quantification tool served from a cloud based web app. Our Minimum Viable Product (MVP) is focused to take input data from our sponsors at Black and Veatch (BV), and create a classification machine learning model that predicts capital expenses across various BV markets.

The data team will write postgres queries from Jupyter-Notebooks with the Python3.5.2 kernel to populate PANDAS dataframes and train machine learning models from the sci-kit learn library and api.

An iterative optimization method will be implemented with the Distributed Evolutionary Algorithms in Python (DEAP) library. Heuristic, iterative optimization methods are ideal to predict merit based on many input variables, when multiple optimal configurations exist. Given a project and its set of unique attributes, what would happen if we changed a subset of those attributes and made another prediction? DEAP can offer suggestions to improve the target metric by answering questions like:

- **Q:** “Let’s change the project location, and see how our prediction changes.”
- **A:** “Our predictive power improves by 50% toward a desirable outcome when the project location is changed from location A to location B.”
Why is the project being undertaken? Describe an opportunity or problem that the project is to address.

In the short term, the project will offer customers insightful, visual, descriptive statistics and predictive analytics for their data. In the long term, the project will securely aggregate data across industries in order to improve the predictive power of all models implemented within the DataDriven community.

What will be the end result of the project?

A web app with “Tableau”-like data visualization technology. A database management system that integrates with various client databases and data types. A prediction pipeline that trains models and offers new insights to customers daily. Each day, DataDriven creates new transformations on previous predictions to inform our customers, who are decision makers and business intelligence leaders.

Project Milestones

- 3-5 specific and measurable objectives per semester for first & second semester

1.) Descriptive statistics and visualizations: Scripts generate target (capex) based descriptive statistics through the lens of each of the input variables. For example, if project size is an input variable for capex, the Python script should generate descriptive statistics for project size vs capex. Time series visualizations can be binned by arbitrary time deltas (annual, quarterly, monthly, etc), covariance matrices and histograms will provide aggregate statistics. Any input variable provided by the customer can serve as the independent parameter.

   Semester one: Draft for BV evaluation and feedback.
   Semester two: Final draft for BV evaluation and feedback.

2.) Predictive analytics and visualization: scikit-learn API is used to train machine learning models and make predictions about future target values based on prior distributions of input parameters. Results are depicted graphically, and supplemented by prior model performance depicted in a confusion matrix.

   Semester one: Draft for BV evaluation and feedback.
   Semester two: Final draft for BV evaluation and feedback.

3.) Genetic algorithms are used to show permutations on input parameters that positively or negatively affect the prediction target. The evolutionary algorithms run nightly to create new insights for the upcoming business day.

   Semester one: Functional GP optimization
   Semester two: Functional GP integration into daily prediction pipeline
• Estimated completion date for each milestone (Draft -> Final Draft)

1.) Dec 15, 2016 -> Feb 15, 2017
2.) Dec 15, 2016 -> March 15, 2017
3.) Nov. 15, 2016 -> April 15, 2017

• Both implementation and documentation milestones

  Documentation for data and web teams will be developed in markdown syntax on the GitLab wiki page. The implementation milestones listed above will also serve as documentation milestones. Eventually, the wiki will be published to serve as the web app’s Technical Support documentation on the public website.

• Gantt Chart:

  We have a pivotal tracker set up to implement Agile methodology for the administrative purposes of this project. Agile methodology is more iterative giving preference to a working model first then touch ups; compared to the Waterfall methodology, where the finished product is the final goal. A gantt chart is suited better for the later but if a rough conversion were to be made to form a gantt chart also submitted as a pdf with this report. But this gantt chart is a very rough estimate and not a good interpretation of our progress plan.

  Documentation for data and web teams will be developed in markdown syntax on the GitLab wiki page. The implementation milestones listed above will also serve as documentation milestones. Eventually, the wiki will be published to serve as the web app’s Technical Support documentation on the public website.
We are using an agile development methodology. The method is organized, prescriptive, and functional. It is not easily converted into a Gannt chart, and conversion to a Gannt chart is redundant. I understand that we need to communicate task ownership and timelines. This is readily available in the GUI at pivotaltraccker.com. You may also view a chart that Quincy created, which contains the same information as a Gantt chart, but it is not visually arranged by date. Please view the chart here. Unfortunately, we have too much data about our progress to encode in this document. We are simply moving too fast. If the chart is not readily digestible, please visit our pivotaltracker profile at https://www.pivotaltracker.com/reports/v2/projects/1863553. All project supervisors should have received invitations to join this group. PivotalTracker is developed to facilitate these types of project management tasks, and it is specifically built for web apps such as ours. Setting it up and learning how to use it was not at all trivial. I hope this will be considered during grading.

If you would like a more detailed report; please feel free to let Quincy or I (Kunal) know so that we can give you a nicer tour through PivotalTracker.

### Project Budget

- **Hardware, software, and/or computing resources**

  All hardware required for this project will be repurposed from personal resources within the group. All software required for this project will have an open source license and will not require a budget.

  Optional memory upgrade for the server: (6x 4GB DDR3 DRAM Memory priced at ~$150)

- **Estimated cost**

  ~$150
• Vendor


• Special training (e.g., VR)

    None

• When they will be required?

    Not required

Work Plan

• Who will do what?

    Omar - Genetic algorithms (TPOT)

    Brian - Algorithms / Genetic Algorithms (DEAP)

    Quinton - Database Specialist (postgres)

    Michael - Statistical Modeling and Prediction (Data Scientist II)

    Kunal - Feature Engineering and Team Lead (Data Scientist I)

Github link: https://datamunger.ddns.net

Preliminary project Design

    The user is greeted with a video explaining our web app services. During playback of
    the video, a small chat window with a picture of a customer services representative asks the
    user: “What are your data analytics needs today?” The answering service is handled by
    Amazon’s Mechanical Turk. Below the video is a login button. The login send new users to
    registration form

    The user registers with the service. An email activation link is sent to the provided email
    address to activate and confirm a valid user. Once registered, the user logs into a
    dashboard. They are immediately greeted with a prompt to specify a prediction task they
    wish for us to solve. They are instructed to provide us with credentials to a database of their
    choice with read only access. If they do not wish to specify this, they may upload csv, xlsx
    format data. Alternatively, they may elect to send us a hard drive at a mailing address to be
    announced at a later date. They submit this form and a new story is created in our
    PivotalTracker workflow, using the PivotalTracker API. Both csv and xlsx format files are
converted used the read_csv or read_xlsx functions from the Python PANDAS API, and saved to our postgres database for analytics tasks.

On registration, the user decides how much data they are willing to share, and a subscription fee is calculated based on the bulk of their sharing. The fee is waived for the duration of a trial prediction task. The user's actions are now complete, they wait for a response from the analysis team.

The analysis team creates a descriptive statistic python notebook, accessible via their DataDriven Dashboard. A simple machine learning model is trained, and accessible via the predictive analytics tab of the dashboard. The predictive analytics tab describes which model was chosen by the analysis team, and why it is an appropriate model for the given task. All source is open and viewable as a Jupyter notebook in html format. This transparency is key, because DataDriven is ultimately a subscription to an anonymized database to an industry vertical. For this minimum viable product (MVP) the industry is production engineering tasks for power industry.

Machine learning predictions improve with the addition of new data. First, we train the model on data provided by the customer. Accuracy, precision, and recall for the target metric are quantified and displayed on screen. In an adjacent panel, the same model is trained on data from the entire industry vertical data set. Accuracy precision and recall are listed for this prediction task. This motivates the value of a subscription to DataDriven, and it provides source code to enable the enterprise customer to begin building their own data science team, or to begin re-training business intelligence personnel on data science programming. Alternatively, we may offer in-house data science experts to operate on an hourly basis.

The user's experience is now locked in. They will decide to continue a subscription, and optionally hire a data scientist to perform additional tasks. Subscription fees are determined by the bulk of data shared with the DataDriven network. All data will be transformed and anonymous. This is the key value of data munging. Enterprise customers may benefit from shared data, without giving away sensitive information about with competitors. Everyone benefits from competitors model training, and DataDriven allows them to do so without risk.

Ethical and Intellectual Property issues

- **Ethical issues**

  Our project's ethical issues are ultimately tied to data sharing. Privacy and anonymity are the key value components of the DataDriven community, and steps are taken to ensure this. We are using an SSL secured server. The SSL certificate is provided by LetsEncrypt.org. The functionality of this certificate may be viewed by visiting [https://datamunger.ddns.net](https://datamunger.ddns.net). The core
ethical concept for DataDriven is to bring the most good to the most people. A subscription costs more for companies who do not share data with us, to encourage what may be considered ethical data sharing. Our responsibilities as the DataDriven team is to be absolutely certain these data are not traceable back to the parent company. Anonymous ID's are used, and transformations on data provide a homogeneous scheme. This task is not trivial, and this is where DataDriven creates value. Once the data are homogenenized and anonymized, they are randomly shuffled so that it is impossible to tell which data belongs to a particular company.

- **IP Issues**

  Our sponsors at BV have been generous to willingly give us all IP. We are using open source methodologies, and so we do NOT take credit for the algorithms we provide. We educate our users on open machine learning toolkits so they may learn how to perform analysis within their own teams. It is not our intention to offer analytics long term, and we do not retain IP rights to the prediction algorithms. This is because we believe in the open source ethos. If everyone has access to machine learning modeling source code, then the softwares will improve as a collaborative effort. Our IP, and our value, is twofold: we will homogenizing the data provided by individual companies so that each customer can develop their own models by querying the relevant columns. Second, we will ensure anonymity. For companies that allow us a persistent database, we will only store metadata that aggregates data from the source tables. This way we avoid storing data on our servers directly, and instead store the functions that transform the data. The customer’s databases will be queried and transformed to comply with our industry vertical shema on demand. This allows us to provide the experience of a single database, without having to store everything in house. This is also a key IP for our project.

**Change Log**

  On the data team, we have a new member, Ryan Feehan, who will be working on genetic algorithms.

On the web team, we have decided to use an architecture that allows our web app to scale. Trailblazer is the architecture of choice.